ДОБАВЛЯЕМ КОНТРОЛЬ ДАННЫХ В ML PIPELINES

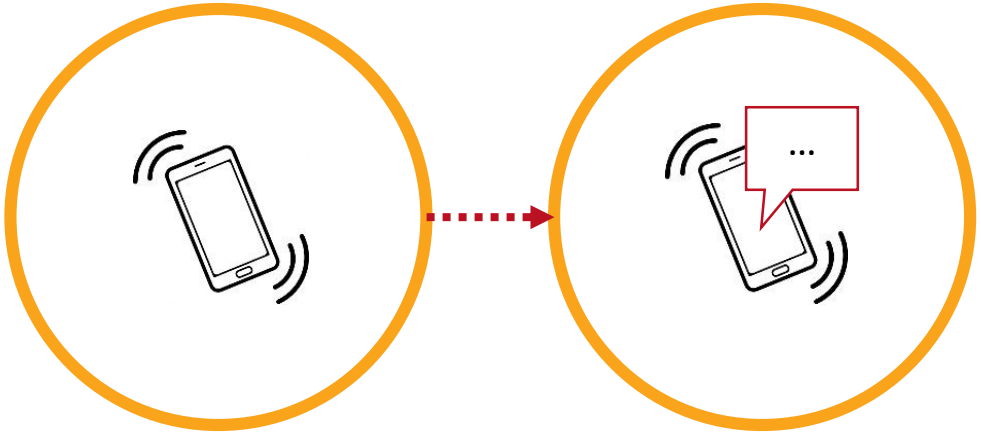
Артем Селезнев (МегаФон)

UseData Conf 2019

Профессиональная конференция для специалистов по машинному обучению и анализу данных

ПРОЕКТ ТРЯСИ СМАРТФОН

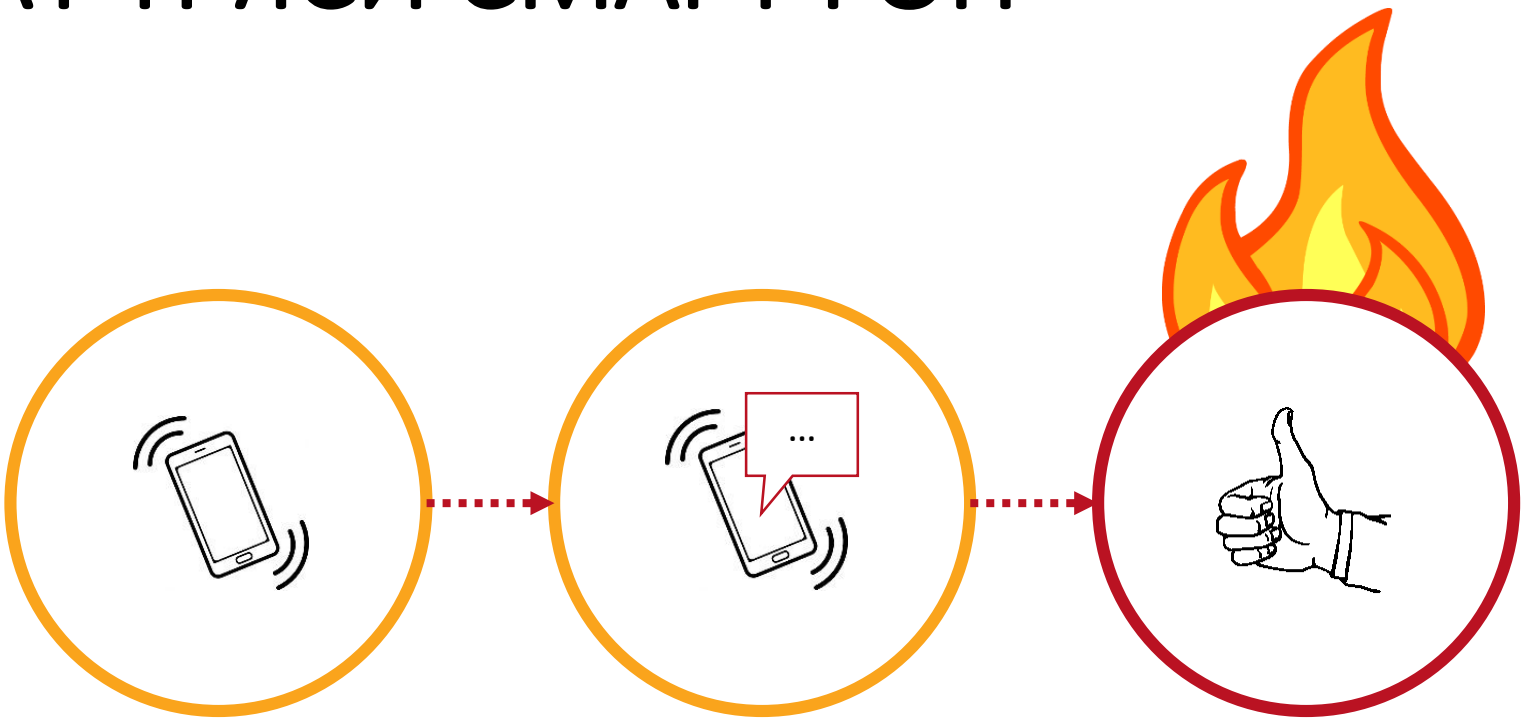
ПРОЕКТ ТРЯСИ СМАРТФОН



ПРОЕКТ ТРЯСИ СМАРТФОН

ПРОЕКТ ТРЯСИ СМАРТФОН

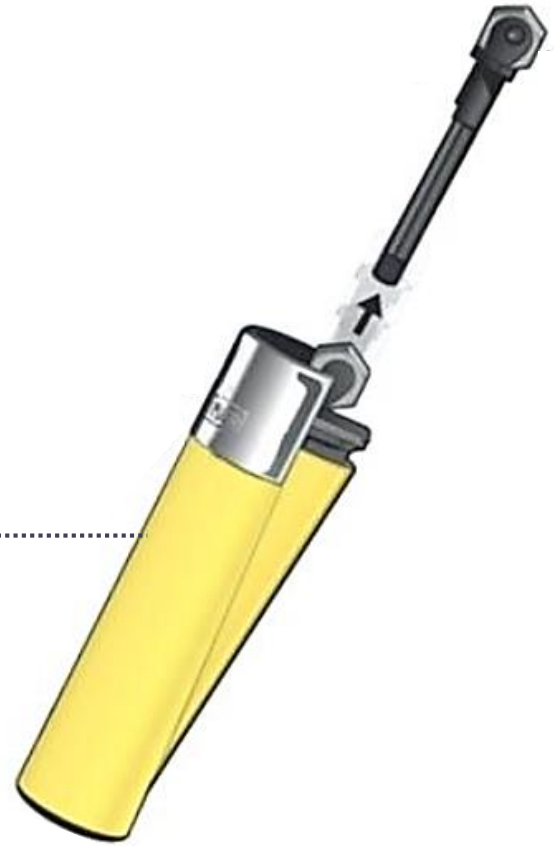
ПРОЕКТ ТРЯСИ СМАРТФОН



НУЖНА «ЗАЖИГАЛКА»

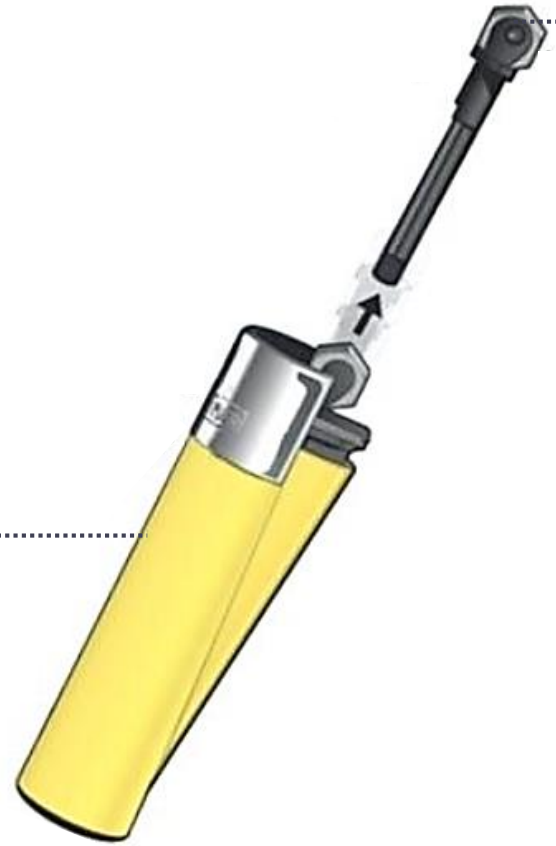
НУЖНА «ЗАЖИГАЛКА»

Данные



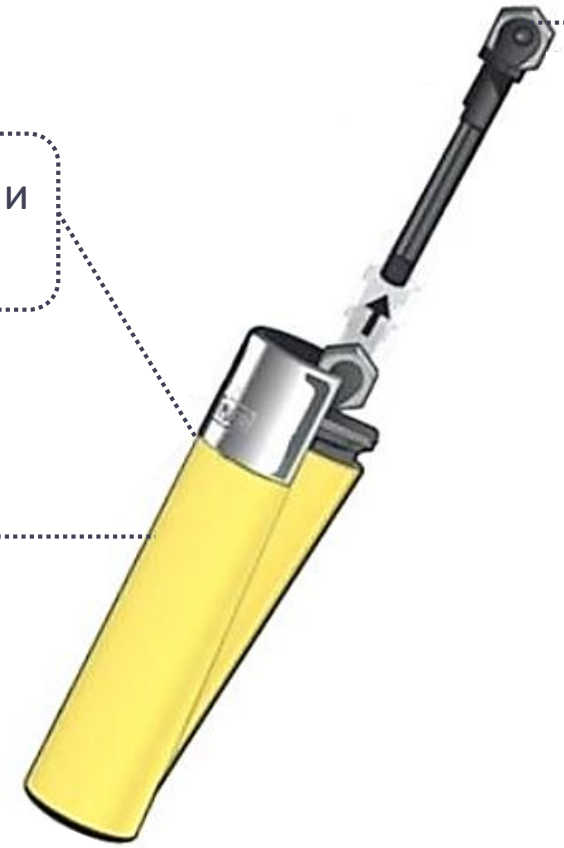
НУЖНА «ЗАЖИГАЛКА»

Заголовки
(headers)



Данные

НУЖНА «ЗАЖИГАЛКА»

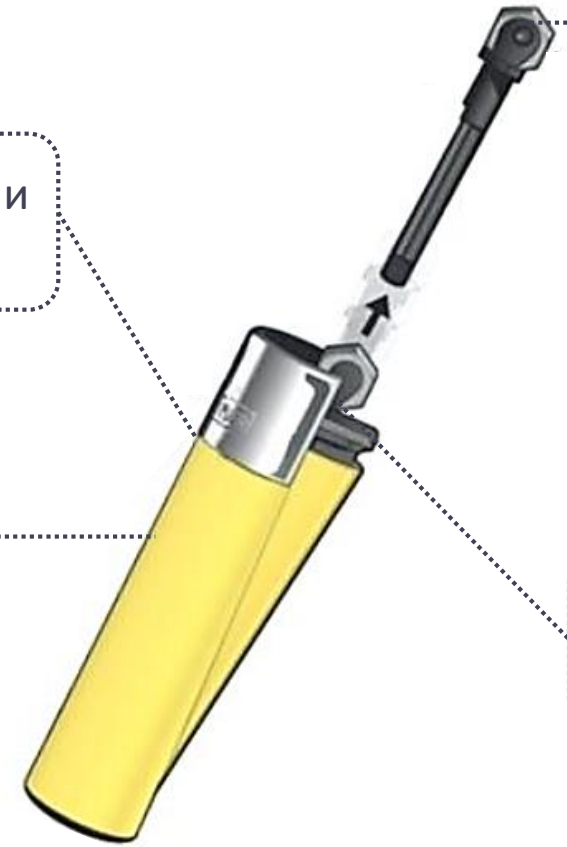


Заголовки (headers)

Обработчики и процессы

Данные

НУЖНА «ЗАЖИГАЛКА»



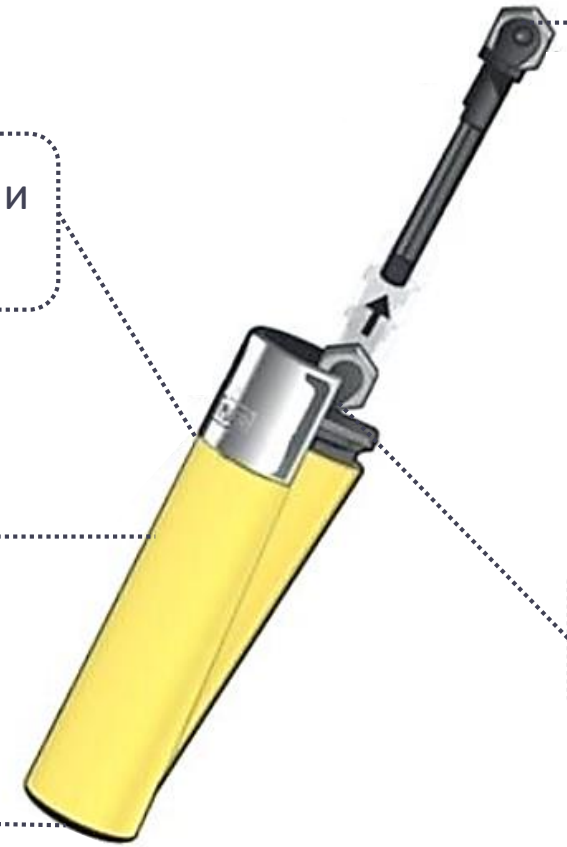
Заголовки (headers)

Обработчики и процессы

Данные

Обученные алгоритмы (pickles)

НУЖНА «ЗАЖИГАЛКА»



Заголовки (headers)

Обработчики и процессы

Данные

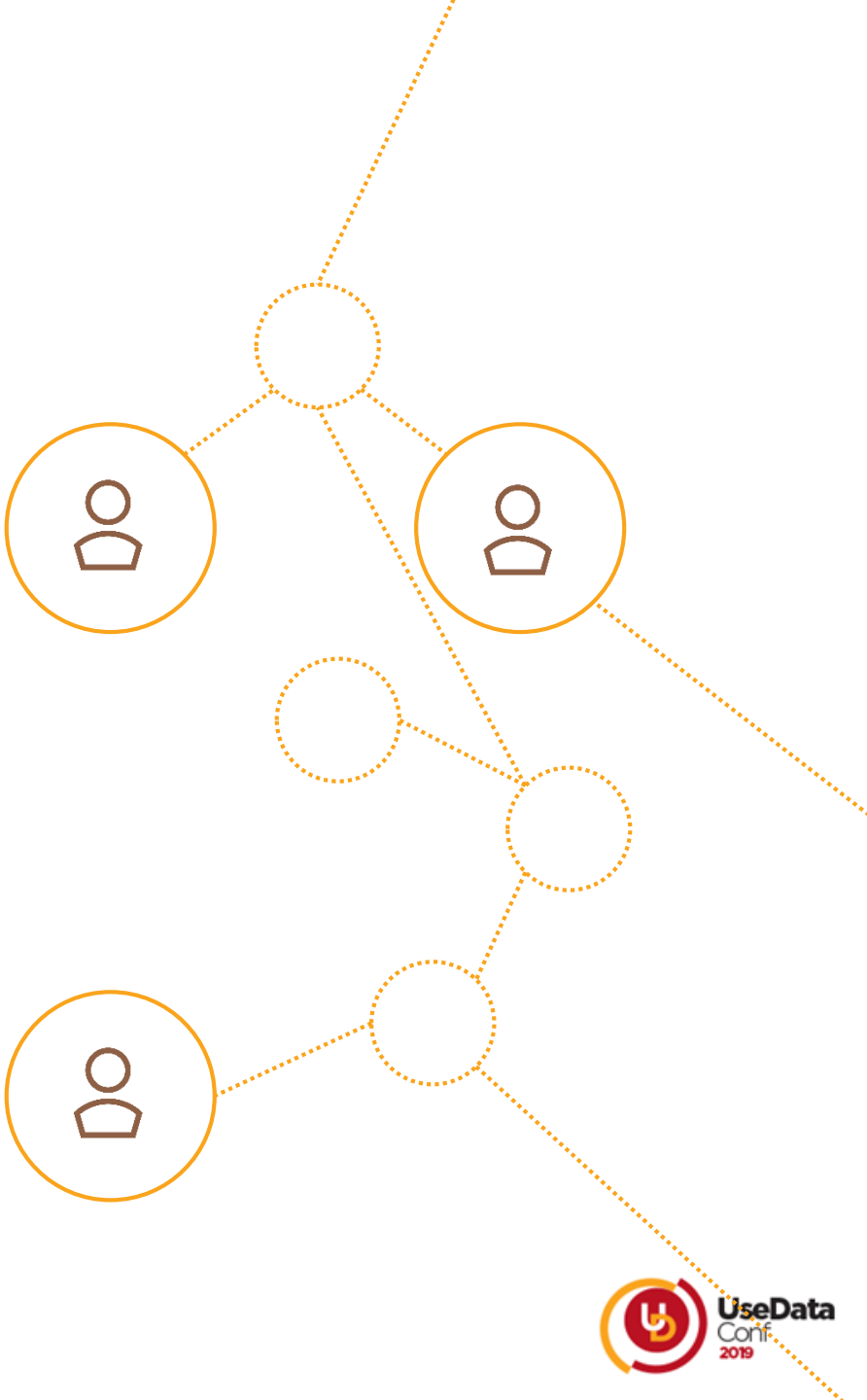
Обученные алгоритмы (pickles)

Предыдущие действия

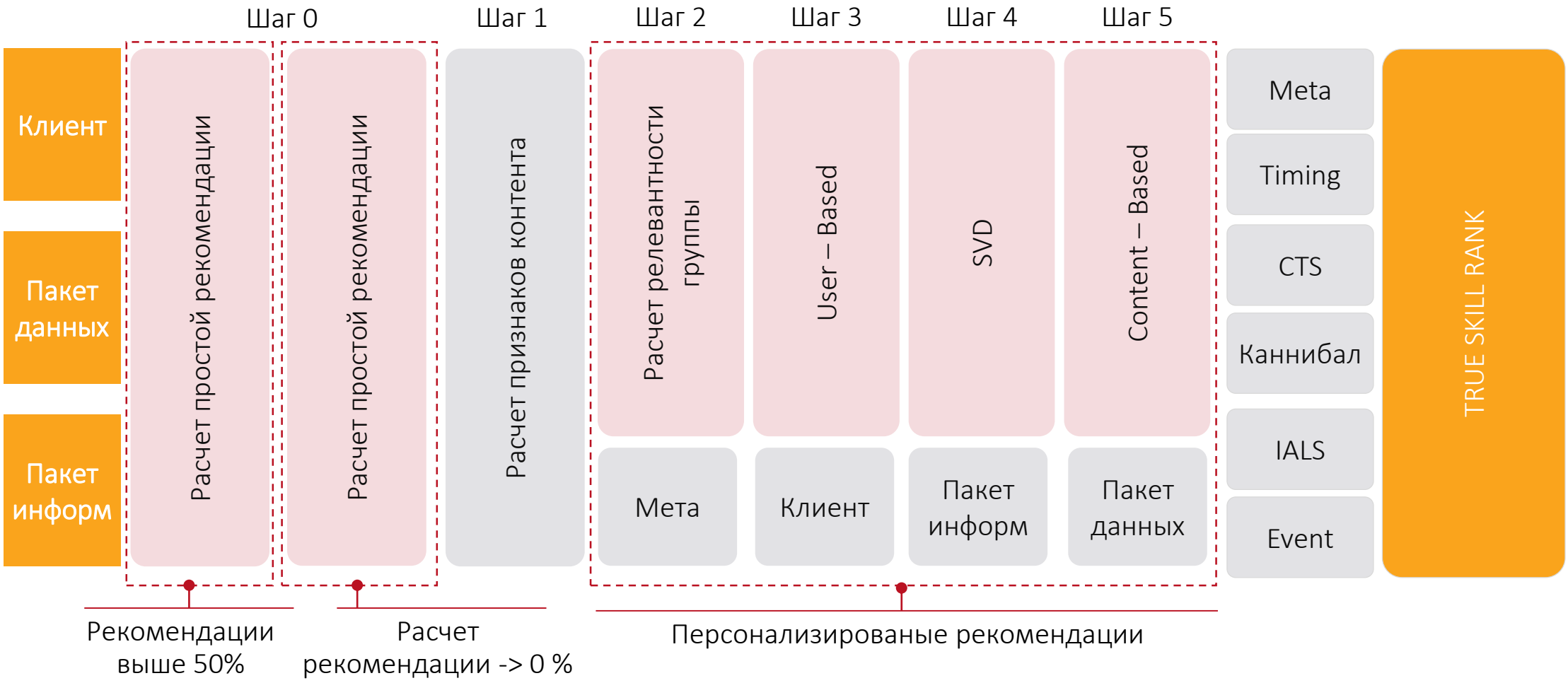
НО ВСЕ МЕНЯЕТСЯ…

НО ВСЕ МЕНЯЕТСЯ…

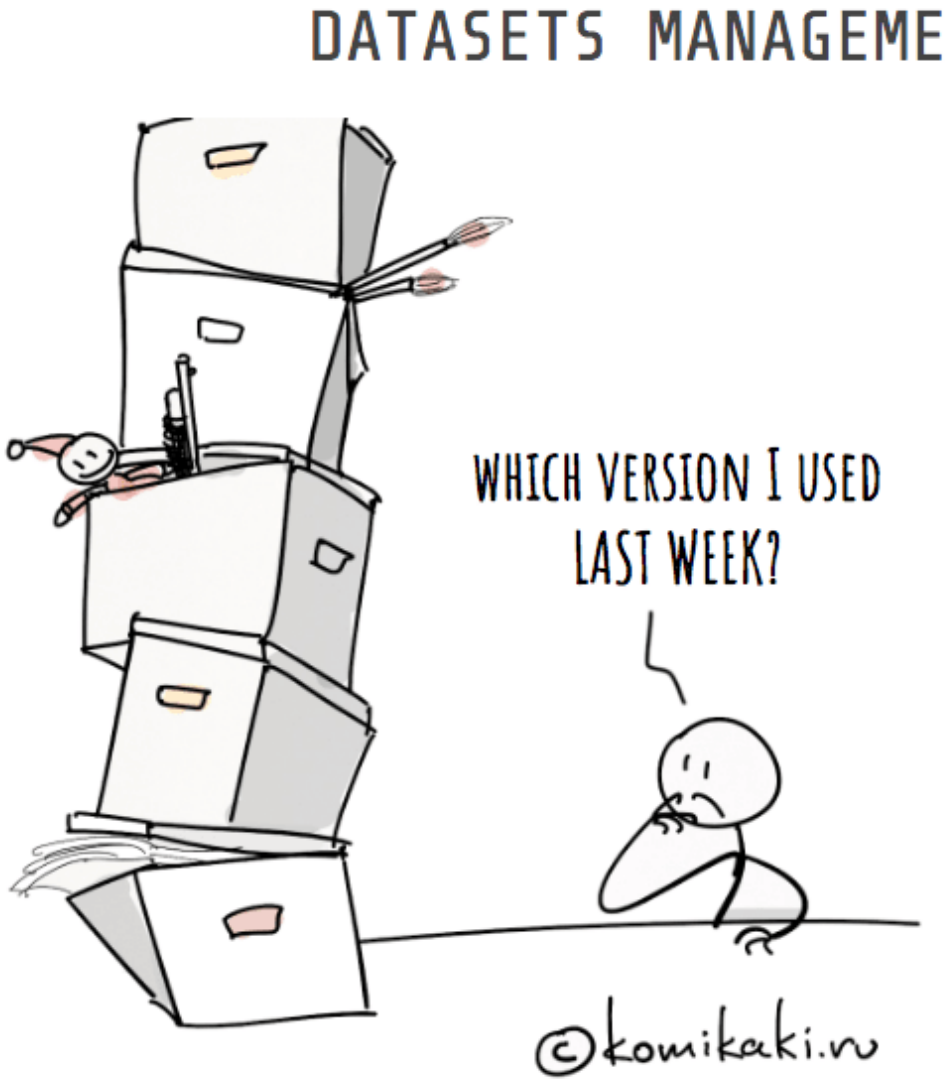
НО ВСЕ МЕНЯЕТСЯ...



У НАС СЛОЖНАЯ «ЗАЖИГАЛКА»



НАСТУПАЕТ КРИЗИС КОНТРОЛЯ ВЕРСИЙ



ПОИСК ПО ТРЕБОВАНИЮ

ПОИСК ПО ТРЕБОВАНИЮ

- Контроль процесса разработки

ПОИСК ПО ТРЕБОВАНИЮ

- Контроль процесса разработки
- Контроль процесса экспериментов

ПОИСК ПО ТРЕБОВАНИЮ

- Контроль процесса разработки
- Контроль процесса экспериментов
- Решение КРИЗИСА КОНТРОЛЯ ВЕРСИЙ

ПОИСК ПО ТРЕБОВАНИЮ

- Контроль процесса разработки
- Контроль процесса экспериментов
- Решение КРИЗИСА КОНТРОЛЯ ВЕРСИЙ
- Воспроизведение экспериментов

ПОИСК ПО ТРЕБОВАНИЮ

- Контроль процесса разработки
- Контроль процесса экспериментов
- Решение КРИЗИСА КОНТРОЛЯ ВЕРСИЙ
- Воспроизведение экспериментов
- Data cache

DVC – РЕШЕНИЕ

- Контроль процесса разработки
- Контроль процесса экспериментов
- Решение КРИЗИСА КОНТРОЛЯ ВЕРСИЙ
- Воспроизведение экспериментов
- Data cache

dvc.org

DVC – ПРОСТОЕ РЕШЕНИЕ

- add
- run
- repro
- push
- …

dvc.org

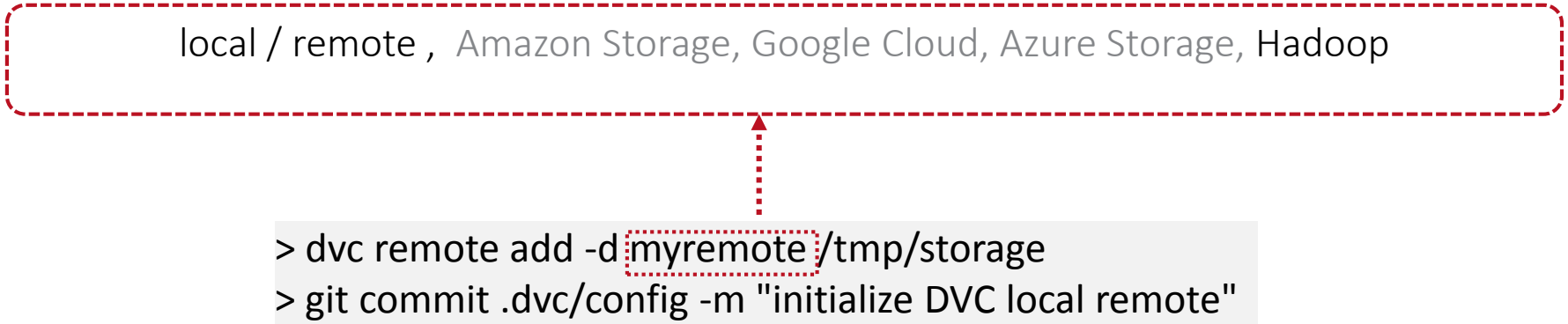
DVC – ПРОСТОЕ РЕШЕНИЕ

- add
- run
- repro
- push
- …

dvc.org

+ GIT

DVC – ПРОСТОЕ ДЛЯ СЛОЖНОГО



DVC – ПРОСТОЕ ДЛЯ СЛОЖНОГО

```
> dvc run  -f catboost_dev/data_pre.dvc \
>              -d constants.csv \
>              -d loader.py \
>              -d preparator.py \
>              -o data/raw_feats.csv \
> python preparator.py
```

DVC – ПРОСТОЕ ДЛЯ СЛОЖНОГО

```
> dvc run  -f catboost_dev/data_pre.dvc \
>          -d constants.csv \
>          -d loader.py \
>          -d preparator.py \
>          -o data/raw_feats.csv \
> python preparator.py
```

DVC – ПРОСТОЕ ДЛЯ СЛОЖНОГО

```
> dvc run -f catboost_dev/data_pre.dvc \
>         -d constants.csv \
>         -d loader.py \
>         -d preparator.py \
>         -o data/raw_feats.csv \
> python preparator.py
```

```
md5: bbcf9272de50…
cmd: preparator.py
wdir: ..
deps:
-   md5: 34rmrfhc58nr…
    path: constants.csv
-   md5: 8a5123fnf11n…
    path: loader.py
-   md5: 90tj58fh4515…
    path: preparator.py

outs:
-   md5: 9a05df6702e9…
    path: data/raw_feats.csv
```

DVC – ПРОСТОЕ ДЛЯ СЛОЖНОГО

```
> dvc run  -f catboost_dev/data_pre.dvc \
>          -d constants.csv \
>          -d loader.py \
>          -d preparator.py \
>          -o data/raw_feats.csv \
> python preparator.py
```

```
md5: bbcf9272de50…
cmd: preparator.py
wdir: ..
deps:
- md5: 34rmrfhc58nr…
  path: constants.csv
- md5: 8a5123fnf11n…
  path: loader.py
- md5: 90tj58fh4515…
  path: preparator.py

outs:
- md5: 9a05df6702e9…
  path: data/raw_feats.csv
```

DVC – ПРОСТОЕ ДЛЯ СЛОЖНОГО

```
> dvc run  -f catboost_dev/.dvc \
>              -d constants.csv \
>              -d loader.py \
>              -d preparator.py \
>              -d data/raw_feats.csv \
>              -d catboost_fitter.py \
>              -m metrics/ctb_binary.txt
> python preparator.py
```

DVC – ПРОСТОЕ ДЛЯ СЛОЖНОГО

```
> dvc run  -f catboost_dev/.dvc \
>          -d constants.csv \
>          -d loader.py \
>          -d preparator.py \
>          -d data/raw_feats.csv \
>          -d catboost_fitter.py \
>          -m metrics/ctb_binary.txt
> python preparator.py
```

```
…

outs:
-    md5: 1f41fedm14dk4…
     path: metrics/ctb_binary.txt
     cache: true
     metric: true
     persist: false
```

DVC – ЧИТАЕМЫЕ ПРОЦЕССЫ

```
> dvc run -f data_pre.dvc \
>         -d constants.csv \
>         -d loader.py \
>         -d preparator.py \
>         -o raw_feats.csv \
> python preparator.py
```

DVC – ЧИТАЕМЫЕ ПРОЦЕССЫ

```
> dvc run -f data_pre.dvc \
>         -d constants.csv \
>         -d loader.py \
>         -d preparator.py \
>         -o raw_feats.csv \
> python preparator.py
```

```
md5: bbcf9272de50…
cmd: preparator.py
wdir: ..
deps:
-   md5: 34rmrfhc58nr…
    path: constants.csv
-   md5: 8a5123fnf11n…
    path: loader.py
-   md5: 90tj58fh4515…
    path: preparator.py

outs:
-   md5: 9a05df6702e9…
    path: raw_feats.csv
```

DVC – ЧИТАЕМЫЕ ПРОЦЕССЫ

```
> dvc run -f data_pre.dvc \
>         -d constants.csv \
>         -d loader.py \
>         -d preparator.py \
>         -o raw_feats.csv \
> python preparator.py
```

```
md5: bbcf9272de50…
cmd: preparator.py
wdir: ..
deps:
-   md5: 34rmrfhc58nr…
    path: constants.csv
-   md5: 8a5123fnf11n…
    path: loader.py
-   md5: 90tj58fh4515…
    path: preparator.py

outs:
-   md5: 9a05df6702e9…
    path: raw_feats.csv
```

```
md5: 8fsdhfe5rceq…
cmd: cat_b.py
wdir: ..
deps:
-   md5: 1ee41dxqecs1…
    path: constants.csv
-   md5: 3x34r45ce14r…
    path: data/raw_feats.csv
-   md5: 8fsdhfe5rceq…
    path: cat_b.py

outs:
-   md5: 15df8tadcacm495…
    path: data/cat_result.csv
```

```
> dvc run -f cat_b.dvc \
>         -d constants.csv \
>         -d raw_feats.csv \
>         -d cat_b.py \
>         -o cat_result.csv \
> python cat_b.py
```

UseData Conf 2019

DVC – ЧИТАЕМЫЕ ПРОЦЕССЫ

```
> dvc run -f data_pre.dvc \
>         -d constants.csv \
>         -d loader.py \
>         -d preparator.py \
>         -o raw_feats.csv \
> python preparator.py
```

```
md5: bbcf9272de50…
cmd: preparator.py
wdir: ..
deps:
-   md5: 34rmrfhc58nr…
    path: constants.csv
-   md5: 8a5123fnf11n…
    path: loader.py
-   md5: 90tj58fh4515…
    path: preparator.py

outs:
-   md5: 9a05df6702e9…
    path: raw_feats.csv
```

```
md5: 8fsdhfe5rceq…
cmd: cat_b.py
wdir: ..
deps:
-   md5: 1ee41dxqecs1…
    path: constants.csv
-   md5: 3x34r45ce14r…
    path: data/raw_feats.csv
-   md5: 8fsdhfe5rceq…
    path: cat_b.py

outs:
-   md5: 15df8tadcacm495…
    path: data/cat_result.csv
```

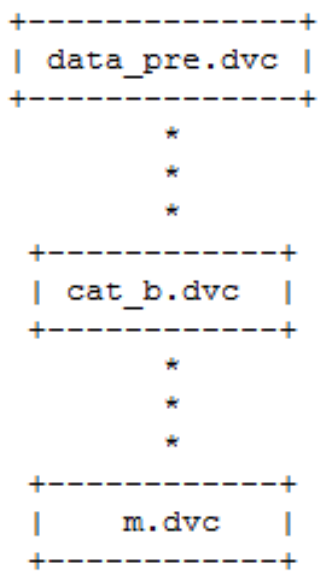
```
> dvc run -f cat_b.dvc \
>         -d constants.csv \
>         -d raw_feats.csv \
>         -d cat_b.py \
>         -o cat_result.csv \
> python cat_b.py
```

```
> dvc run -f m.dvc \
>         -d cat_resutlt.csv \
>         -d cat_eval.py \
>         -m tb_m.txt \
> python cat_eval.py
```

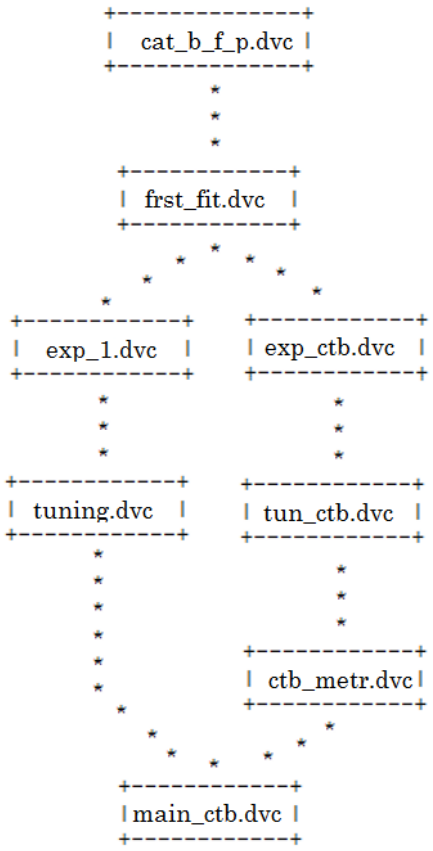
```
md5: 93miew4123cdas…
cmd: cat_eval.py
wdir: ..
deps:
- md5: 15df8tadcacm495…
  path: data/cat_result.csv
- md5: 8fsdhfe5rceq…
  path: cat_b.py

outs:
- md5: 4t63tcm5t23x4…
  path: metrics/ctb_m.txt
  cache: true
  metric: true
  persist: false
```

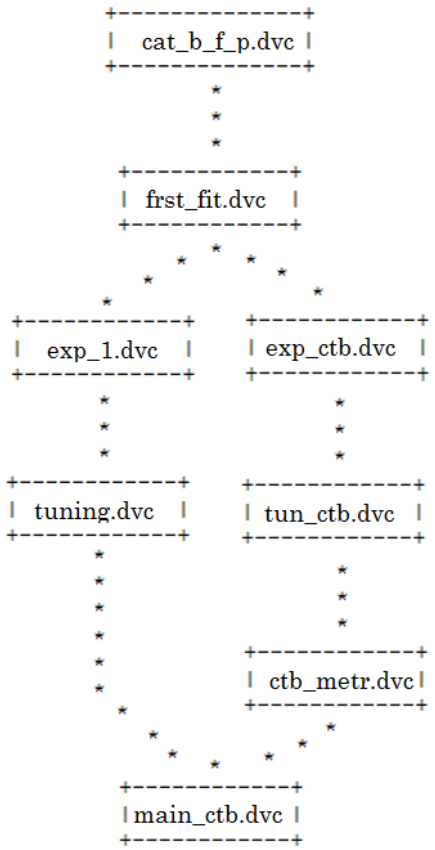
НЕ НРАВИЛАСЬ ВИЗУАЛИЗАЦИЯ

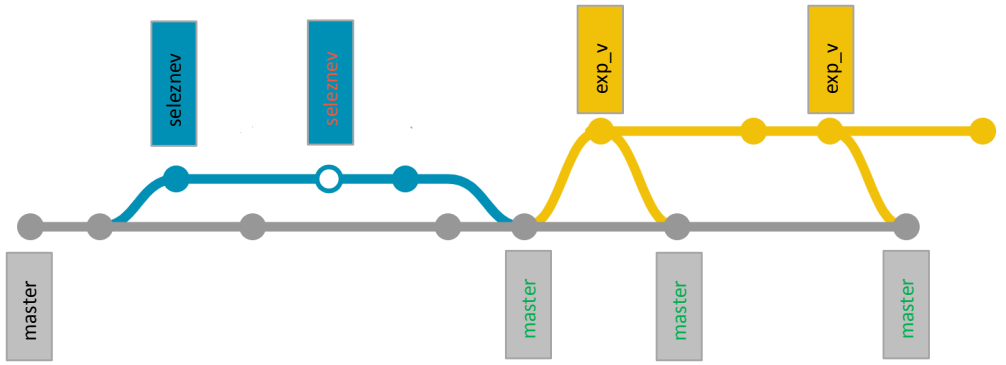


НЕ НРАВИЛАСЬ ВИЗУАЛИЗАЦИЯ



НЕ НРАВИЛАСЬ ВИЗУАЛИЗАЦИЯ



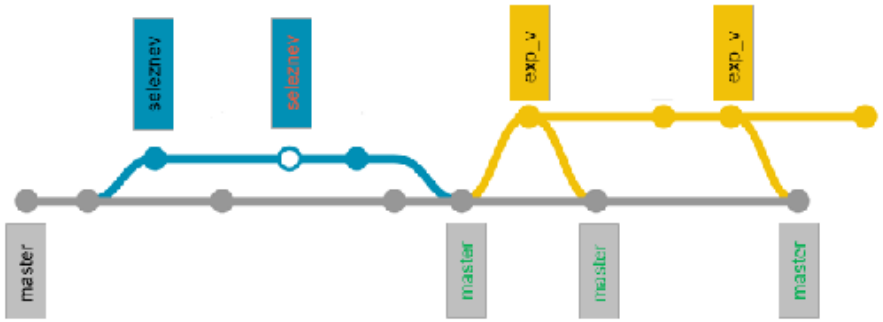


ШАГ В АВТОМАТИЗАЦИЮ

```python
import dvc

with open('controller.dvc', commits = 5, 'auto') as d:
    d.display()
```

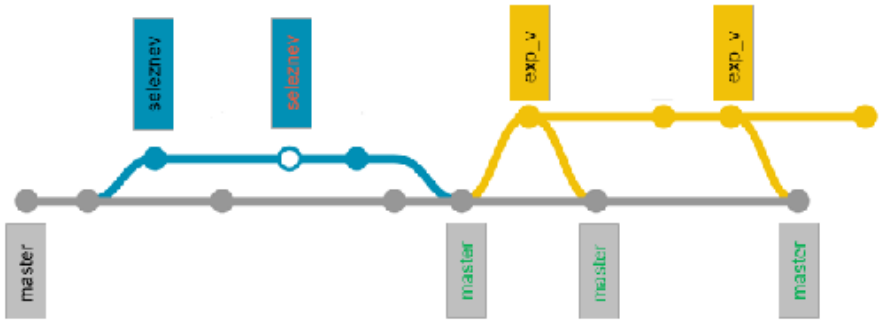
```
Last change 20:17 11.08.2019, branch name - Seleznev
Last status - True, Metrics - True, Metrics - AUC_ROC .67
```



ШАГ В АВТОМАТИЗАЦИЮ

```
1  import dvc
2
3  with open('controller.dvc', commits = 5, 'auto') as d:
4      d.display()
```

```
Last change 20:17 11.08.2019, branch name - Seleznev
Last status - True, Metrics - True, Metrics - AUC_ROC .67
```

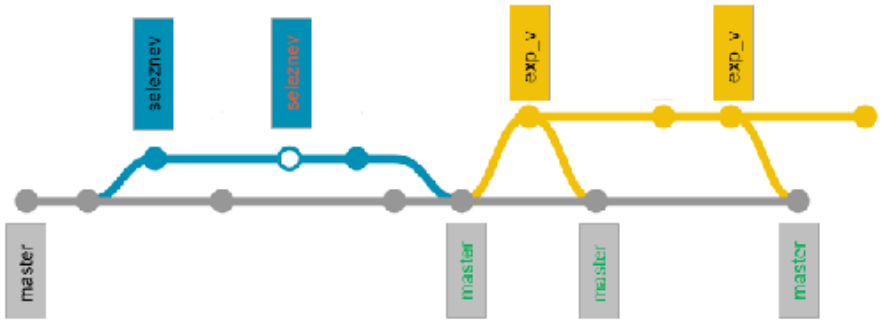


- Контекстный менеджер

ШАГ В АВТОМАТИЗАЦИЮ

```
1  import dvc
2
3  with open('controller.dvc', commits = 5, 'auto') as d:
4      d.display()
```

```
Last change 20:17 11.08.2019, branch name - Seleznev
Last status - True, Metrics - True, Metrics - AUC_ROC .67
```



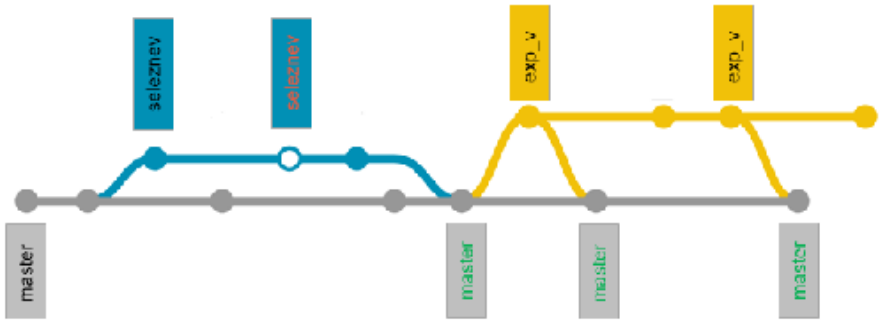
- Контекстный менеджер
- Автокоммиты (шт/час)

ШАГ В АВТОМАТИЗАЦИЮ

```python
import dvc

with open('controller.dvc', commits = 5, 'auto') as d:
    d.display()
```

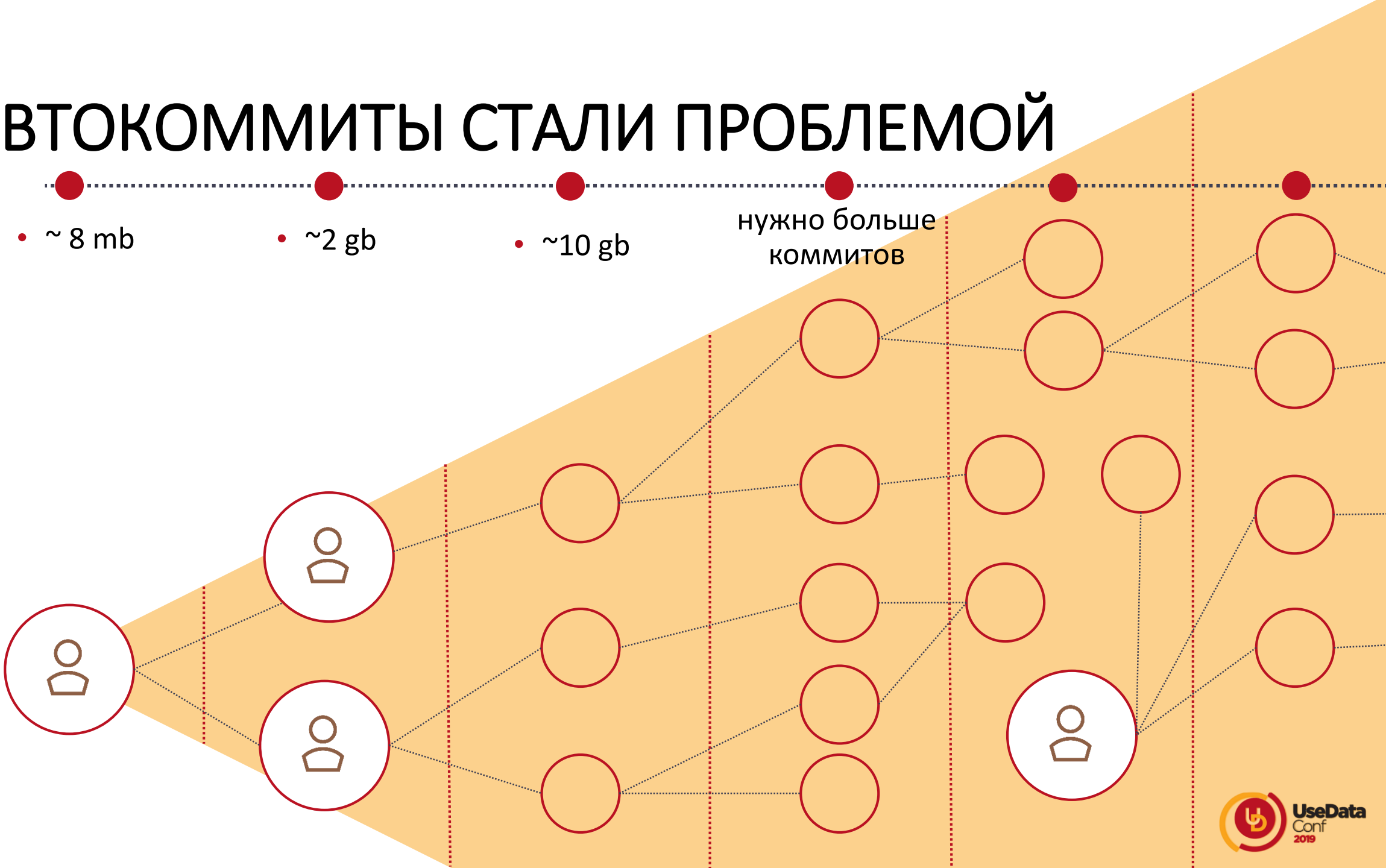
```
Last change 20:17 11.08.2019, branch name - Seleznev
Last status - True, Metrics - True, Metrics - AUC_ROC .67
```



- Контекстный менеджер
- Автокоммиты (шт/час)
- Редактирование pipeline

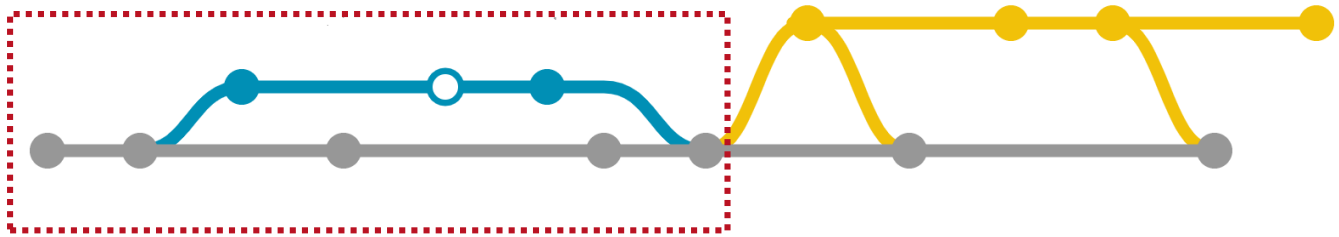
АВТОКОММИТЫ СТАЛИ ПРОБЛЕМОЙ

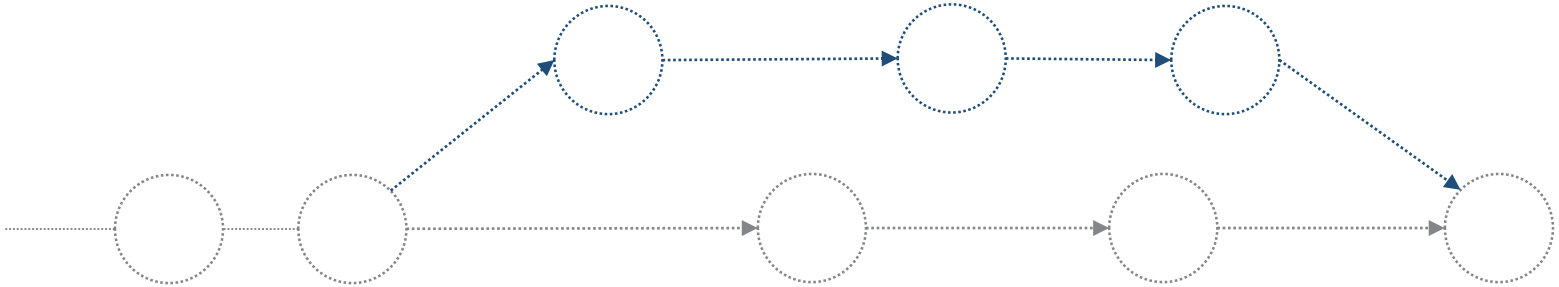
- ~ 8 mb
- ~2 gb
- ~10 gb
- нужно больше коммитов



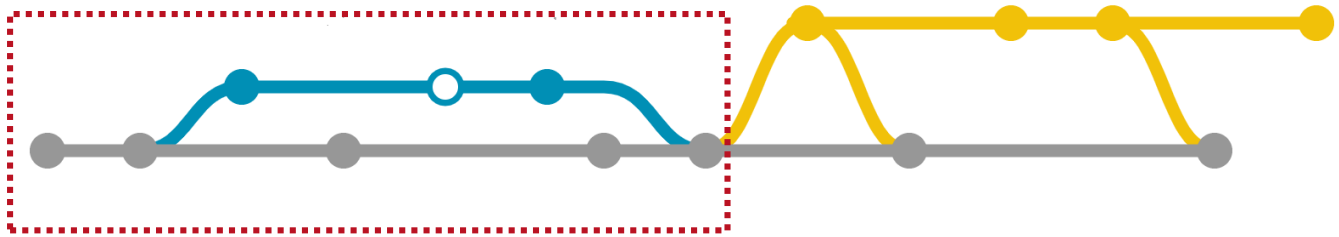
СЕТИ ПЕТРИ – РЕШЕНИЕ

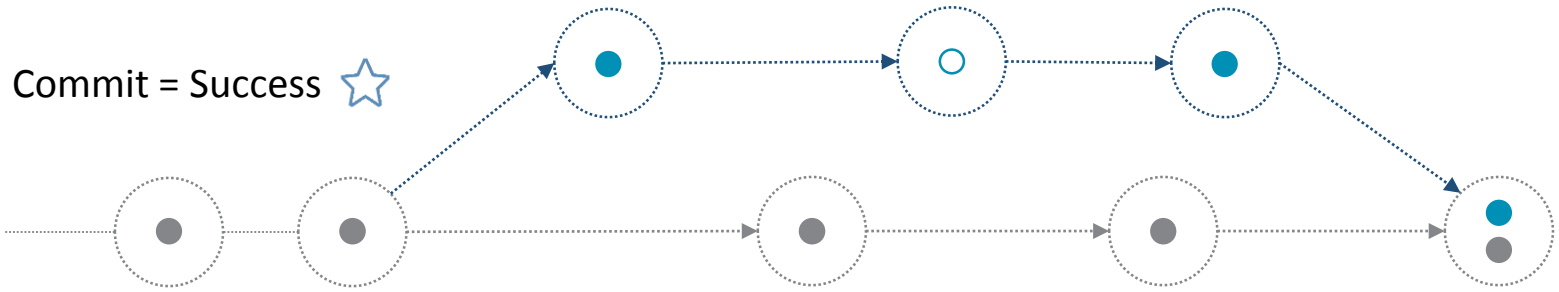
СЕТИ ПЕТРИ – РЕШЕНИЕ



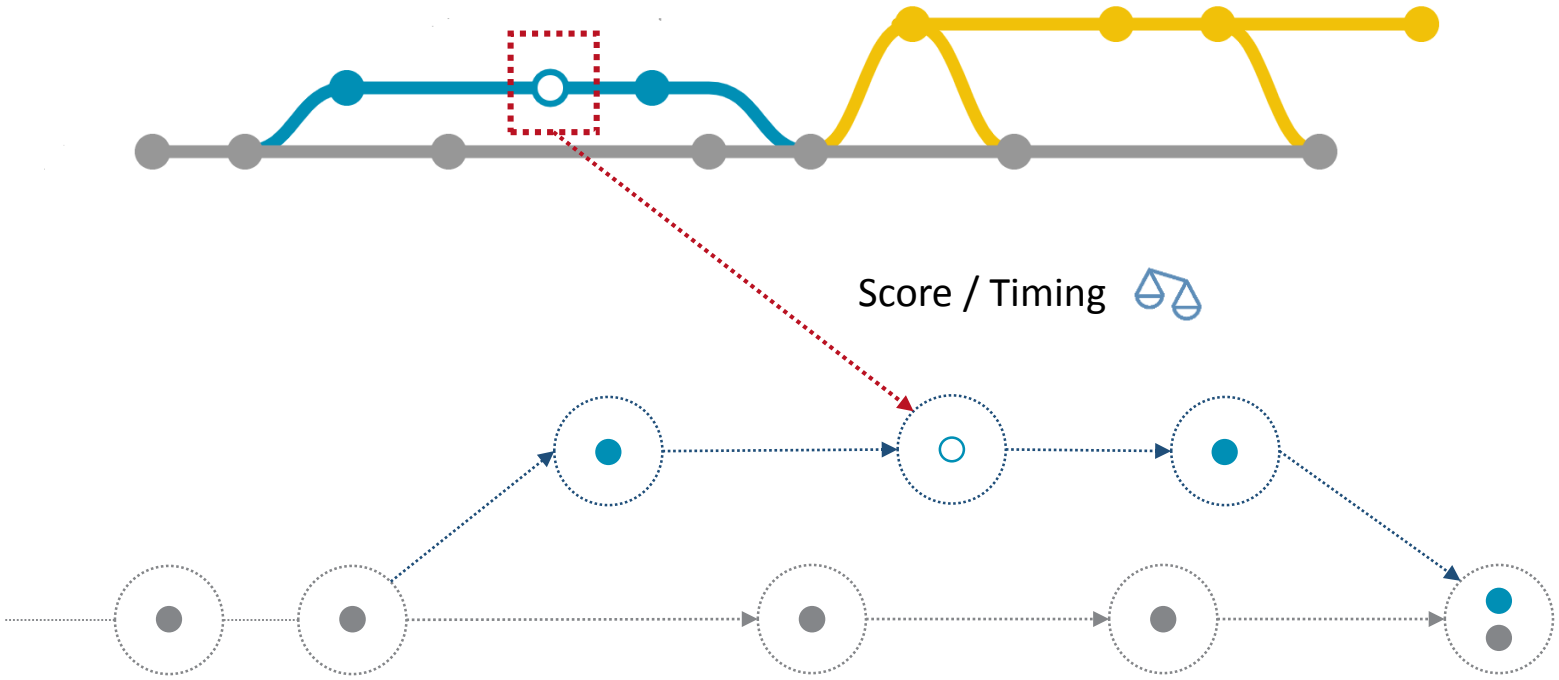


СЕТИ ПЕТРИ – КОНТРОЛЬ КОММИТОВ

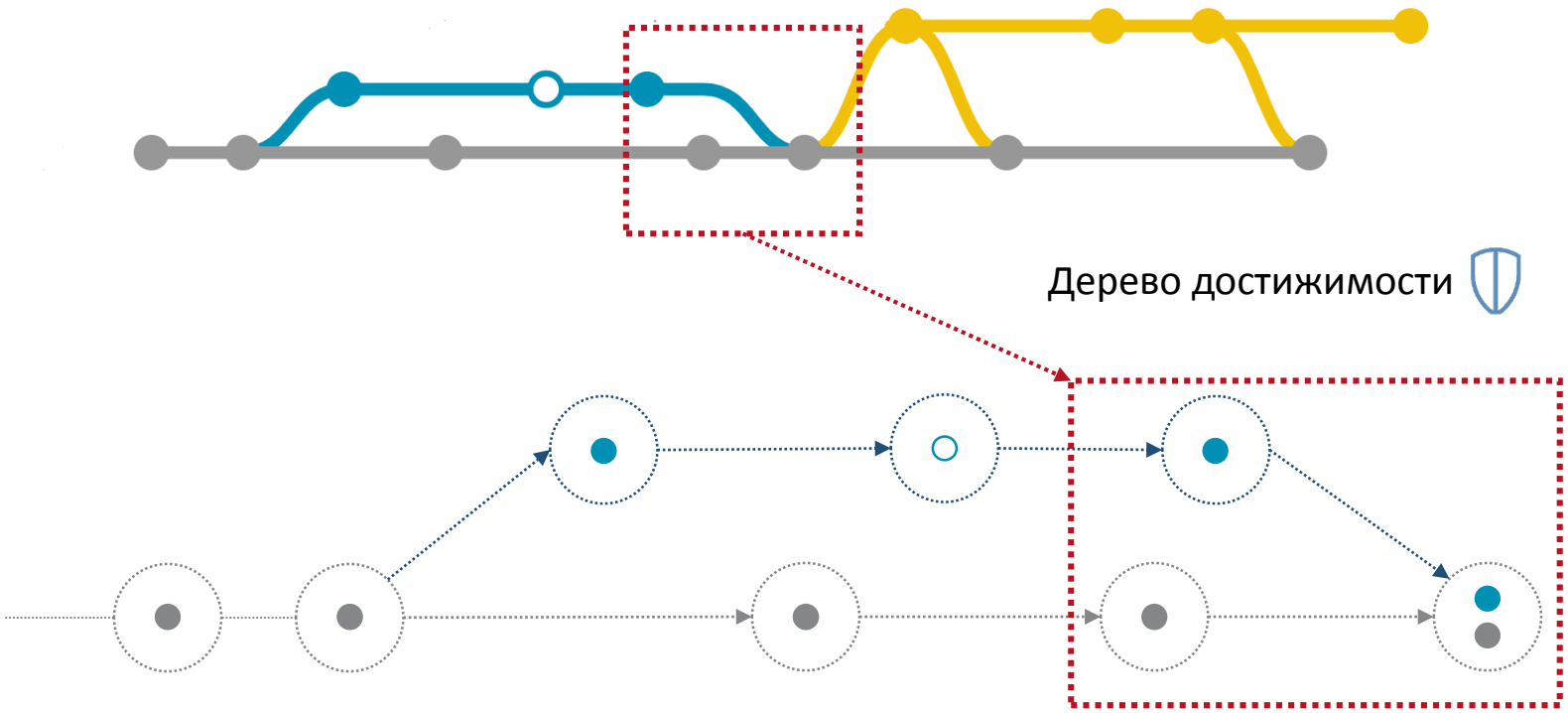




СЕТИ ПЕТРИ – СОРЕВНОВАНИЕ РЕЗУЛЬТАТОВ



СЕТИ ПЕТРИ – АВТОСЛИЯНИЕ



DVC РЕШИЛ ПРОБЛЕМЫ

dvc.org

- Контроль дата-сетов для модели
- Контроль сопутствующей информации
- Воспроизводство экспериментов
- Контроль процесса экспериментов

ДОБАВЛЯЕМ КОНТРОЛЬ ДАННЫХ В ML PIPELINES

Артем Селезнев (МегаФон)

@SeleznevArtem
facebook.com/seleznev.artem.info

UseData Conf 2019

Профессиональная конференция для специалистов по машинному обучению и анализу данных